Non-local Intrinsic Decomposition with Near-infrared Priors

Ziang Cheng
Australian National University
Australia
ziang.cheng@anu.edu.au

Yinqiang Zheng
National Institute of Informatics
Japan
yqzheng@nii.ac.jp

Shaodi You
Data61, CSIRO
Australia
shaodi.you@data61.csiro.au

Imari Sato
National Institute of Informatics
Japan
imarik@nii.ac.jp

Abstract

Intrinsic image decomposition is a highly under-constrained problem that has been extensively studied by computer vision researchers. Previous methods impose additional constraints by exploiting either empirical or data-driven priors. In this paper, we revisit intrinsic image decomposition with the aid of near-infrared (NIR) imagery. We show that NIR band is considerably less sensitive to textures and can be exploited to reduce ambiguity caused by reflectance variation, promoting a simple yet powerful prior for shading smoothness. With this observation, we formulate intrinsic decomposition as an energy minimisation problem. Unlike existing methods, our energy formulation decouples reflectance and shading estimation, into a convex local shading component based on NIR-RGB image pair, and a reflectance component that encourages reflectance homogeneity both locally and globally. We further show the minimisation process can be approximated by a series of multi-dimensional convolutions, each within linear time complexity. To validate the proposed algorithm, a NIR-RGB dataset is captured over real-world objects, where our NIR-assisted approach demonstrates superiority over RGB methods.

1. Introduction

Recognised as a fundamental problem for scene understanding by computer vision research [4, 18, 24, 27, 44], intrinsic decomposition is the task of recovering from the input image two logically independent factors - a shading image that is solely dependent on illumination, shadows and shapes, and a reflectance image that is solely dependent on textures and colours. This independency is nullified once the input image is observed, permitting decomposition albeit under an extremely underdetermined system. In this paper, we address this ambiguity by a novel NIR-assisted algorithm that combines both reflectance and shading priors in a unified probabilistic model.

Due to its application background and ill-posed nature, single image intrinsic decomposition has attracted extensive interest in computer vision research, prompting a variety of RGB-based approaches via traditional priors [45, 21, 41, 42, 36, 6] or data-driven regression models (*e.g.* deep learning) [34, 3, 23, 43, 17, 32]. However, the information provided by an RGB image alone is limited. A common problem is weak textures and strong shadows are often misclassified. More recently, further advances have been made into the hyper-spectral domain where additional subspace constraints can be sought [10, 22]. Still, these methods ultimately work within the visible spectrum, and a practical concern of hyper-spectral methods is the difficulty of image acquisition, which requires a dedicated and usually slow hyper-spectral camera that precludes any scene motion.

We propose an algorithm that solves for intrinsics with the help of near infrared (NIR) imagery. The NIR band has a longer wavelength than visible lights and reveals fairly different scattering patterns. The uniqueness of NIR band leads to its success in assisting visibility enhancement [39, 33], surface reconstruction [16, 15] and semantic segmentation [38, 14]. In this paper, we exploit the empirical observation that texture variation is considerably reduced in this channel. This property of NIR lessens the ambiguity of decomposition and offers a powerful prior for shading. On the other hand, synchronised RGB-NIR cameras are being popularised by vendors and are available on the market (*e.g.* dual-CCD camera, or single sensor camera with RGB-NIR 4-channel Bayer pattern), making our set-up more cost-friendly and practical compared to general hyper-spectral methods.

We approach NIR-RGB intrinsics via gradient-based energy minimisation. Since the NIR image has less texture variation but preserves shape and illumination conditions, we exploit it for an informative and straightforward shading smoothness prior that is robust to texture variations. Additionally, a non-local, dense energy term is designed for the reflectance homogeneity assumption, which allows our method to recover reflectance across long geometric distances and pass disjoint reflectance regions. We show this energy term can be re-written as a convolution, thus making it efficiently solvable. To validate our approach, we collect NIR-RGB image pairs of various objects, where our algorithm demonstrates robustness against both textures and shading variations and compares favourably against RGB-based algorithms. Our main contributions are highlighted as follows.

- We advance intrinsic decomposition by exploiting NIR band for a simple yet powerful shading prior.
- Departing from other non-local methods, we develop a dense but linear-time-differentiable energy formulation that can be minimised directly by a gradient-based optimiser. Our algorithm involves no $O(n^2)$-complex clustering/grouping process thus scales better to image size.
- We capture an object-level NIR-RGB dataset with partly known ground truth intrinsics to facilitate future research.

2. Related work

The RGB-based intrinsic decomposition gained popularity following a multichannel adaptation of Retinex theory [28, 19, 26, 21]. Various papers have suggested that RGB-based methods suffer significantly less ambiguity than their grayscale counterparts, due to the introduction of chromaticity features that reside in a null space of shading variations [21, 3]. It is further discovered that chromaticity values are often drawn from several basis colours (sparsity). This observation leads to non-local reflectance constraints [41, 42, 36, 20, 47] that cluster/group pixels based on their chromatic or texture affinity.

Some methods rely on additional inputs for guidance. Common methods from this category include user-assisted (scribbles) algorithm for image [8] and video [7] intrinsics. Chen *et al*. [12] proposed an RGB-D algorithm that further decomposes shading image into direct and indirect irradiance components. Barron *et al*. [3] developed a data-driven algorithm that jointly estimates object shape, illumination and reflectance and extended it to scene level images with the additional input from depth sensors [2]. Huang *et al*. [22] used multi-spectral images by constraining intrinsic components in a low dimensional subspace along the spectral domain.

Recent advances in synthetic datasets [9, 11, 30] and weakly-supervised/unsupervised training routines enabled many deep-learned approaches. Kim *et al*. [25] trained a neural network to jointly estimate depth and intrinsic components by minimising a CRF energy (loss) function. Shi *et al*. [43] and Li *et al*. [31] proposed large-scale synthetic datasets for training. Janner *et al*. [23] trained different subnets to predict individual intrinsic components, from which a differentiable rendering algorithm is built to allow self-supervision from reconstruction loss. Zhou *et al*. [48] trained a model to learn the ordering of reflectance pixels and incorporate this prior in CRF energy. Fan *et al*. [17] applied a flexible loss layer for training a universal model on both fully-labeled and weakly-labeled datasets. Several self-supervised approaches [31, 32, 29] rely on multiple images with varying illumination conditions to train a network without ground truth by enforcing identical reflectance; Yu *et al*. [46] extended this idea and used a multi-view pipeline to recover scene geometry. LapPyrNet [13] is a multi-scale network that predicts each layer in the Laplacian pyramid of intrinsics. Baslamisli *et al*. [5] proposed (a) a physically-based image formation loss and (b) a network that separates image gradients following Retinex theory.

3. Methodology

In this paper, we opt for a traditional energy minimisation scheme in the hope to develop a theoretical, controllable approach that can work without any training data.

Our energy consists of three components for shading, reflectance, and posterior regularisation, respectively. In the following sections, we shall first introduce the overall energy formulation then detail each component individually.

3.1. Intrinsic decomposition as energy minimisation

By assuming a linear response camera model, the image I can be written as the sum over a diffuse reflection I_d and a specular reflection I_s

$$I = I_d + I_s, \tag{1}$$

where the diffuse reflection is assumed here to be dominant over specular reflection (*i.e*. $I_d \gg I_s$) and can be further factorised into a shading component S and a reflectance component R. Hence

$$I \approx I_d = S \times R. \tag{2}$$

With I observed, we want to solve for the most likely shading image S and reflectance image R that describe I. In this paper, we formulate this problem in a probabilistic framework by minimising the following energy

$$E(S, R) = E_S(S) + E_R(R) + E_I(S, R), \tag{3}$$

where E_S and E_R are the shading and reflectance likelihood terms respectively, and E_I is a posterior term that permits the presence of *e.g*. image noise and non-diffuse reflection.

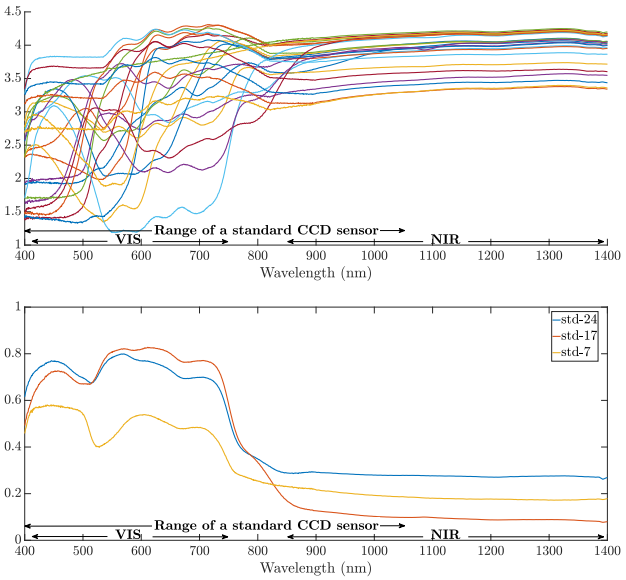


Figure 1: Top: reflectance r versus wavelength curves for 24 colourants on a colour checker. Bottom: Standard deviation of r along colourant domain - a subset of 17 (out of 24) colourants have close reflectance values in NIR range, while the other 7 have values spread more apart but still closer in NIR than in visible bands.

In this paper, we shall assume illumination is white light.[1] For notational simplicity, we will use big letters I, S and R for input/output image pixel values, and small letters i, s and r for their corresponding log values.

3.2. NIR-assisted shading energy E_S

Research showed that NIR band is transparent to a range of colourants/dyes [37]. This observation leads the NIR image to be considered as a direct surrogate for shading over some unique materials [40, 15]. However, the generality of these material-dependent assumptions is limited, as numerous colourants (*e.g.* Carbon black, which reflects no light in a wide spectrum) are still visible in NIR band.

To further demonstrate the usefulness and limitation of NIR band, multi-spectral reflectance values are sampled over 24 colourants on a colour checker, as illustrated in figure 1. Note, that the spectral curves become gradually flattened out and aggregated as wavelength increases. As a result, a group of 17 colourants have much less reflectance variance in NIR range than in visible range (in which case NIR image approximates shading). However, the other 7 still differ in reflectance even under NIR band (though variation is still much less than in visible range), in which case the NIR image fails to represent shading. See Fig. 2 for an example of how different colorants appear under NIR.

[1] For rank-1 illumination, this can be achieved by a white balancing step (see our experiments).

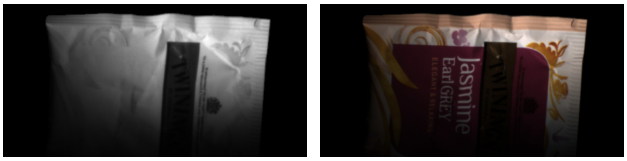


Figure 2: A visual comparison between NIR (left) and RGB (right) images of a tea bag. Note, that while texture variation is significantly reduced in the NIR image, some patterns (*e.g.* flowers and texts) are still visible due to the use of black pigments for colouring purposes. Also, while compared to RGB image, the upper part of NIR image appears brighter but the lower slightly darker.

Additionally, we found in our experiments that sometimes NIR images cannot faithfully represent shading, even on uniform material. This is because some real-world materials exhibit (slightly) different reflective properties under NIR and visible bands. An example is the tea bag in Fig. 2 - when compared to RGB image, the log-contrast between the upper and lower part of the same material (white, uncoloured plastic) under NIR band is greater. This suggests material's bidirectional reflectance distribution function (BRDF) could indeed be wavelength-dependent, and a NIR-as-shading assumption would lead to inconsistent RGB reflectance estimation.

Based on these observations, we intend to develop a NIR prior for shading that generalises well to a wide range of dyes/materials. It turns out that we can do so by simply penalising local shading variation where it is greater than that of NIR image

$$E_S^{\text{nir}}(s) = \sum_{x,y \in \mathcal{N}} \Big(\max(0, |s_x - s_y| - |i_x^{\text{nir}} - i_y^{\text{nir}}|) \Big)^2, \quad (4)$$

where $\mathcal{N}$ denotes a neighboring pixel set, and i_x^{nir} is the log brightness of the NIR image at pixel x.

Intuitively, this energy term acts as a soft version of hard constraint $|i_x - i_y| \leq |i_x^{\text{nir}} - i_y^{\text{nir}}|$, which dictates local variation of shading image should not exceed that of NIR image. This is an arguably reliable constraint based on the observation in Fig. 1, and provides a tight bound on image regions where the reflectance component is barely visible in NIR spectrum. Surprisingly, this straightforward formulation plays a significant part in reducing shading ambiguity caused by textures and works well in our experiments.

It is worth noting, that minimising E_S^{nir} alone leads to a trivial solution space (*e.g.* one such minimum is at $s_x \equiv s_y \ \forall x, y$). To remedy this, we blend E_S^{nir} with a NIR-

modified Colour Retinex [21] energy, defined as

$$E_S^{\text{rtx}}(s) = \sum_{x,y\in\mathcal{N}} (s_x - s_y - \epsilon_{xy})^2 \quad s.t. \tag{5}$$

$$\epsilon_{xy} = \begin{cases} G(i_x) - G(i_y) & \text{if } \|C(i_x) - C(i_y)\|_2 < T^C \text{ and} \\ & |G(i_x) - G(i_y)| < T^G \\ i_x^{\text{nir}} - i_y^{\text{nir}} & \text{otherwise if } |i_x^{\text{nir}} - i_y^{\text{nir}}| < T^N \\ 0 & \text{otherwise} \end{cases},$$

where $C(i_x)$ and $G(i_x)$ are functions that return the RGB chromaticity and log brightness of pixel x, respectively, and T^C, T^G and T^N are thresholding parameters (T^C is fixed at 0.1 as the optimal value on MIT dataset [21]). Compared with vanilla Colour Retinex, a major departure of Eq. (5) is that when an RGB-based reflectance edge occurs, it may use the gradients of NIR log-image to guide shading estimation, instead of simply letting shading to be locally constant. Eq. (5) reduces to plain Colour Retinex when $T^N = 0$. For more details on Colour Retinex, we refer the reader to [21].

The overall shading energy thus becomes

$$E_S(s) = E_S^{\text{nir}}(s) + E_S^{\text{rtx}}(s). \tag{6}$$

This formulation is convex and first-derivative-continuous, making it globally minimisable via gradient-based methods.

3.3. Non-local reflectance energy E_R

Our reflectance energy combines a local component E_R^{loc} and a non-local component E_R^{nol}

$$E_R(r) = (1-\alpha)E_R^{\text{loc}}(r) + \alpha E_R^{\text{nol}}(r), \tag{7}$$

where α is a weighting factor fixed at 0.9 to down-weight the semi-dense global energy. The purpose for reflectance energy is to encourage reflectance homogeneity (*i.e.* flattened reflectance image) both locally and globally.

The local term E_R^{loc} is defined as

$$E_R^{\text{loc}} = \sum_{x,y\in\mathcal{N}} w_{xy} \sum_{c\in\text{RGB}} (r_x^c - r_y^c)^2 \quad where \tag{8}$$

$$w_{xy} = \begin{cases} 1 & \text{if } \|C(i_x) - C(i_y)\|_2 < T^C \text{ and} \\ & |G(i_x) - G(i_y)| < T^G \\ & \text{, or} \|I_x^{\text{RGB}} - I_y^{\text{RGB}}\|_2 < \epsilon \\ 0 & \text{otherwise} \end{cases}.$$

Eq. (8) is partially similar to Eq. (5), in that it encourages local reflectance consistency while adjacent pixels share similar values (in this paper, we use a small $\epsilon = 5e-3$ for normalised image). However, Eq. (8) breaks energy transfer between two pixels when there is a reflectance edge (by letting $w_{xy} = 0$), and works on RGB channels instead of gray scale.

Previous research has shown that reflectance values are often drawn from several 'basis colours', and thus reside within some low-dimensional subspace/manifold of colour system. However, this belief cannot be modeled by local term E_R^{loc} (or any path-based algorithm in general, *e.g.* Retinex), because the message chain is blocked wherever there is a reflectance edge. Common methods in this direction typically utilize sparsity constraints by clustering pixels by their colour or texture affinity and minimising the intra-cluster distances [42, 36, 6]. Here we propose a novel non-local energy term based on convolution. Our energy formulation differs from existing methods, in that:

- There is no clustering involved. Clustering algorithms generally have quadratic time complexity thus limited scalability, while our energy term can be calculated and differentiated within linear time.
- We encourage reflectance homogeneity by reducing some distance measure between reflectance values, instead of making an explicit sparsity assumption. This allows reflectance values to vary gradually, instead of appearing distinct or overspreaded in colour space.

We start by giving the mathematical formulation for non-local term

$$E_R^{\text{nol}}(r) = \sum_{x,y} \sum_{c\in\text{RGB}} \frac{k(f_x, f_y)}{\sum_z k(f_x, f_z)} (r_x^c - r_y^c)^2 \tag{9}$$

$$= \sum_{x} \sum_{c\in\text{RGB}} \left(r_x^{c\,2} - 2r_x^c \frac{\mathcal{K}_x * r_y^c}{\mathcal{K}_x * 1} + \frac{\mathcal{K}_x * r_y^{c\,2}}{\mathcal{K}_x * 1} \right), \tag{10}$$

where $k(f_x, f_y)$ is some affinity measure (kernel function) between two feature vectors f_x and f_y (which we shall discuss later). The purpose for normalization term $\sum_z k(f_x, f_z)$ is to avoid bias by any dominant colour. In Eq. (10) we use the convolution operator $\mathcal{K}_x*$ to denote a convolution centred at f_x with kernel function $k(f_x, \cdot)$ (*i.e.* $\mathcal{K}_x * v_y = \sum_y k(f_x, f_y) v_y$ for any vector-valued v_y). With this convention, Eq. (9) is reduced to a convolution at coordinates $\{f_x\}$ in some feature space. Note, that Eq. (9) models an edge between every two pixels in reflectance image (a total of $O(n^2)$ edges, where n is the number of pixels), thus its computation becomes quickly intractable for a mid-sized image. However, by re-writing it into Eq. (10), we shall see that it can be computed in $O(n)$ time.

In practice, we employ an iterative gradient-based solver for minimising energy. For t^{th} iteration, the feature vector $f_x^{(t)}$ is defined as

$$f_x^{(t)} = [w_X u_x \; w_X v_x \; w_C C(r_x^{(t-1)})^T \; G(r_x^{(t-1)})]^T, \tag{11}$$

where $w_X = 500$ and $w_C = 5$ are constant weights, $C(\cdot)$ and $G(\cdot)$ are chromaticity and brightness functions respectively, and (u_x, v_x) are the image coordinates of pixel x.

The choice for kernel function plays a critical part in the performance of our algorithm. Here we project feature

points $\{f_x\}$ onto a manifold called permutohedral lattice [1] for convolution. Permutohedral lattice is intended for a fast approximation to bilateral filtering, in the form of

$$k(f_x, f_y) \approx \begin{cases} e^{-|f_x - f_y|^2/2\sigma^2} & \|f_x - f_y\|_\infty < \lambda \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where λ and σ are two constant values when dimension of f_x is fixed. This truncated form has the advantage to disconnect energy transfer between two pixels if their colour difference or geometric distance exceeds a threshold, in which case we no longer force them to be similar. On the other hand, if two pixels have similar values and are not far away on image plane, they are encouraged to take similar values. As a result, sharp edges or distinct colours (likely caused by reflectance change) of the input image will be preserved in reflectance estimation, while gradual changes (even across long geometric distances or reflectance edges, which Retinex cannot handle) will be flattened out. The rationale behind this behaviour is similar to that of clustering. However, a full convolution on permutohedral lattice can be performed within linear time complexity w.r.t. number of pixels, which makes the calculation and differentiation of Eq. (10) scalable. For the exact algorithm of permutohedral lattice, we refer the reader to [1].

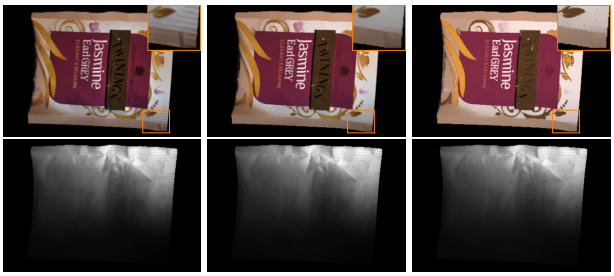
An interesting side note is with Eq. (10) and Eq. (11), computation and differentiation of Eq. (9) is conceptually similar to three passes of bilateral filter (on $\mathbf{1}$, r and r^2 each) - an image processing technique that has been shown to improve reflectance estimations [35].

3.4. Posterior term E_I

Most existing methods assume pure diffuse (Lambertian) shading and impose the hard constraint $i = s + r$ to reduce the number of variables. However, this has several limitations. For example, an inaccurate (often over-smoothed) shading estimation can cause some shading components to 'bleed into' reflectance image. Another example is the dark image regions that are susceptible to camera noise - since this greatly affect chromaticity, artifacts may appear on reflectance image even when shading estimation is precise. For these reasons we relax $i = s + r$ and solve for s and r as two variables, whose dependency is modeled by a posterior term that allows the presence of some level of uncertainties (caused by *e.g.* noise, over-exposure, non-grey-scale light sources, or non-Lambertian reflectance components *etc.*)

$$E_I(s, r) = \beta\big(\big(I - \exp(s+r)\big)^2 + 0.05\big(i - s - r\big)^2\big). \tag{13}$$

The first term accounts for a Gaussian distributed error on observed pixel values, while the second term (with much smaller weight) primarily acts as a quadratic regularizor of the log search space that helps to overshoot local minima (*e.g.* while $exp(s + r) \to 0$, gradients of first term vanish, in which case we rely on second term for minimising Eq. (13)). By adjusting the weighting parameter β, we control how strictly $i = s+r$ is satisfied, as illustrated in Fig. 3.



(a) $\beta = 100$ (b) $\beta = 5$ (c) $\beta = 1$

Figure 3: Reflectance and shading estimations using different values of β. Typically, large β values cause some shading component to 'leak into' reflectance image and noise visible in underlit regions (zoom in for details), when too small β results in an over-flattened reflectance image.

4. Optimisation process

The full energy model consists of both convex and non-convex components, all of which are linear-time differentiable. Considering many components are near quadratic, we employ an L-BFGS algorithm as our energy minimiser. The full optimisation process is described in Algorithm 1.

Algorithm 1: Iterative MAP solver

$s \leftarrow 0$;
$r \leftarrow i$;
while *not converged* **do**
  $f_x \leftarrow$ Eq. (11);
  $s, r \leftarrow$ one iteration of L-BFGS on $E(s, r)$;
return $S = exp(s)$, $R = exp(r)$

A side note is that since all energy components are linear-time-differentiable, one could plug them into the loss function of a CNN as a means to self-supervision while preserving the scalability of training.

5. Experiments

5.1. NIR-RGB image collection

Our dataset contains object-level images. We use a linear-response RGB camera without NIR cut filter to collect image pairs - a 400-700nm VIS pass filter and a 850nm long pass filter are used for capturing RGB and NIR images respectively. The filters are mounted on a motor-driven filter wheel that is placed in front of the camera, which ensures the camera position is fixed despite different filters

being used. A xenon lamp is used for illumination purpose, and wide-band/wide-grid polarizers are placed on both light source and camera to suppress specular lights. Our camera is equipped with a 12-bit ADC, and raw images are saved in 16-bit format. Foreground object masks are obtained by manual annotation.

The camera is dark-calibrated by taking control images of an unlit dark room. Pixel values in these images are uniform due to forced cooling and are subtracted from all RGB-NIR images. Images of a standard white target are also captured for white-balancing purpose. The above pre-processing steps are applied for all methods presented in following sections for fair comparison.

We adopt the methods in [21] for obtaining the ground truth intrinsics. To acquire ground truth shading component, we remove reflectance by either spraying a thin layer of white coating on colourful objects, or by photographing white objects before we colour it. We call these two subsets **testsetA** and **testsetB**, repectively. In the latter case, different paints are used to add reflectance components to both RGB and NIR images.

5.2. Comparison with RGB-based method

Due to the scale of dataset, we cannot re-train data-driven methods on it. For this reason we compare our algorithm with two state-of-the-art methods trained on similar object-level datasets, *i.e.* SIRFS [3] and ShapeNet-pretrained CNN [43], to minimize dataset bias. Colour Retinex [21] is also included as a baseline method for comparison and is trained on entire testsets.

We fix β at 2 during quantitative evaluation. With the remaining two paratemers T^G and T^N, we enforce the constraint $T^N = \max(0, T^G - 0.5)$ and thereby reduce the number of actual free parameters to 1 to avoid overfitting. We run a line search of T^G in $\{0.7, 0.9, 1.1, 1.3, 1.5\}$ using cross-validation, and choose the best performing values for both **testsetA** and **testsetB**. Results are shown in Table 2, where we use LMSE and scale-insensitive MSE metrics for fine-grained and global evaluation[2]. Some sample images from both testsets (two from each) are illustrated in the top half of Fig. 4.

	shading		reflectance		avg. score	
	LMSE	MSE	LMSE	MSE	LMSE	MSE
SIRFS[3]	54.1	**44.9**	91.1	39.1	72.6	**42.0**
ShNet[43]	51.3	72.7	75.5	36.7	63.4	54.7
CR[21]	201	676	85.9	152	144	414
Ours	**47.9**	**46.7**	**58.9**	**34.2**	**53.4**	**40.5**

Table 1: Quantitative results on **testsetA**. Best performing methods (*i.e.* within 5% of lowest score) are bolded.

[2]Values are scaled up by a factor of 1000

	shading		reflectance		avg.	
	LMSE	MSE	LMSE	MSE	LMSE	MSE
SIRFS[3]	9.16	9.44	8.25	16.4	9.30	12.3
ShNet[43]	**5.43**	**4.01**	26.5	29.7	15.97	16.9
CR[21]	7.10	4.60	7.89	**11.0**	7.50	**7.80**
Ours	5.80	8.26	**4.25**	**11.0**	**7.03**	**7.64**

Table 2: Quantitative results on **testsetB**. Best performing methods (*i.e.* within 5% of lowest score) are bolded.

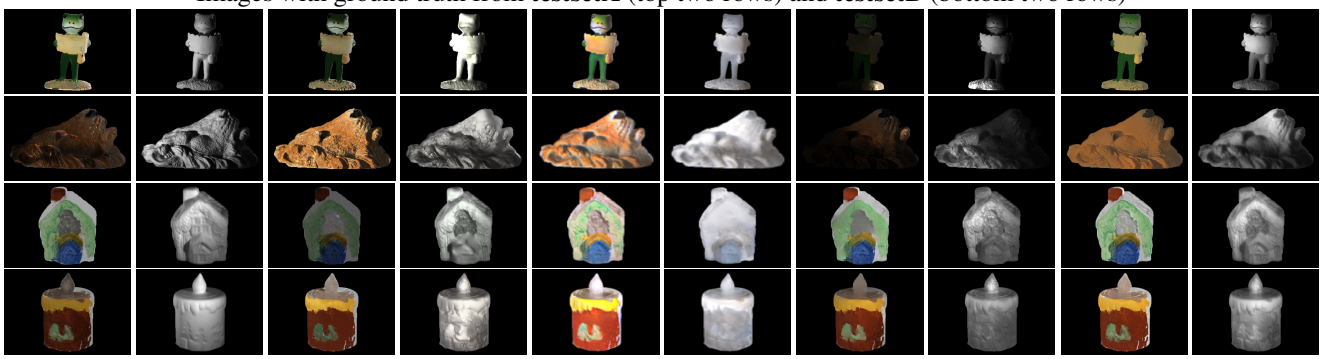
It can be observed that all methods produce significantly worse results on **testsetA** than on **testsetB**. This is because **testsetB** is created from painting white objects with a few distinct colour patches, and therefore have much less reflectance variation than **testsetA**. Surprisingly, the plain Color Retinex on average performs better than more sophisticated RGB methods. An interesting observation is that while ShapeNet-CNN yields the best shading prediction, its reflectance estimation is by far the most inaccurate, which suggests the network may have a preference for certain colours over others (see Fig. 4 for evidence). While compared with Color Retinex, our algorithm gives better reflectance estimation. One possible reason is the global reflectance homogeneity assumption is well-satisfied by **testsetB**.

TestsetA contains much more challenging and ambiguous cases, with a mixture of weak textures and strong shadows. Particularly, some images have reflectance components that resemble shades, and shading components that appear to be textures. Our NIR-assisted energy obtains the best performance on this testset, as the NIR channel is much less hindered by the pseudo-shading texture components.

We also perform a qualitative study for a variety of real-world objects. By allowing β to vary between $[1, 10]$ (2 free parameters), a visual comparison is given in the bottom half of Fig. 4. Notably, data driven methods (*i.e.* SIRFS[3] and ShapeNet[43]) yield less desirable results, suggesting they may have overfitted training dataset. While SIRFS yields visually appealing results, its shading estimation is not always accurate. More specifically, texture sometimes interferes with shading estimation, and prediction is also susceptible to irregular, concave surfaces. Both SIRFS and ShapeNet cannot handle dark shades, in which case shading is often overestimated.

By comparing RGB images with their NIR counterparts, we observe that texture variance is either reduced or completely disappear in NIR image. Once again, this confirm the usefulness of NIR spectrum in assisting shading estimation. On the other hand, our global reflectance energy is able to overcome strong shades and enforce colour homogeneity across long distance and reflectance edges. Overall, our method arguably produces the best visual results on average.

Images with ground truth from **testsetA** (top two rows) and **testsetB** (bottom two rows)



(a) Ground Truths (b) SIRFS[3] (c) ShapeNet CNN[43] (d) Colour Retinex[21] (e) Proposed method

Images w/o ground truth

(f) RGB/INR inputs (g) SIRFS[3] (h) ShapeNet CNN[43] (i) Colour Retinex[21] (j) Proposed method

Figure 4: Visual results on our dataset. Data-driven methods are placed on the left-hand side and handcrafted methods are placed on the right. Images with ground truth are given in the top and those without in the bottom. All images are normalised to $[0, 1]$ for display. See supplementary materials for results.

5.3. Ablation study and variations of energy form

We conduct a quantitative ablation study on **testsetA** to validate the effectiveness of global reflectance energy as well as that of NIR-based priors, by considering several special cases of our energy formulation shown in table3. An **RGB-local** algorithm arrives where $w_N = 0$, $\alpha = 1$ and a uniform image is given in place of NIR input. This reduces the energy formulation to one closely resembling Colour Retinex, albeit with the additional posterior term (*i.e.* relaxed energy). Surprisingly, this relaxed version performs

	input		energy		avg	
	RGB	NIR	local	global	LMSE	MSE
RGB local	✓		✓		76.2	87.7
RGB global	✓		✓	✓	62.5	57.0
NIR local	✓	✓	✓		58.4	52.2
full	✓	✓	✓	✓	53.4	40.5

Table 3: Several special cases used in ablation study.

much better than vanilla Color Retinex. Some possible reasons are (1) we use a different optimiser for solving energy, with a more strict convergence criterion than implementation of [21], (2) compared to [21], we use a window of 3 pixels instead of 1, and (3) our formulation is more robust than the inversely proportional constraint $I = S \times R$ (*e.g.* when S is small, an erroneous estimation of S could severely impact R, and vice versa).

RGB global is yet another special case where $\alpha \neq 1$, but both w_N and T_N are 0. With the addition of global reflectance energy, we see a significant improvement on both LMSE and MSE metrics. Table 4 compares **RGB global** with the sparsity constraint [36] over the results on **testsetA**. It is seen that **RGB global** has a better score on LMSE

	avg. score	
	LMSE	MSE
RGB global	62.5	57.0
Gehler *et al.* [36]	80.0	52.0

Table 4: Comparison between **RGB global** and classic sparsity constraints [36].

but lags slightly behind in MSE, which suggests Gehler *et al.* [36]'s method is more capable of restoring overall reflectance but fails to capture fine-scale details. On the other hand, solving **RGB global** proves to be much more efficient than the iterative clustering process [36] (a few minutes versus up to an hour, depending on different images).

NIR local is a local method with $\alpha = 1$ and the NIR input. The NIR input leads to a greater improvement than **RGB global** over **RGB-local**. At this stage, the algorithm already outperforms [43] on both metrics. To further demonstrate the usefulness of global reflectance energy, a visual comparison is given in Fig. 5, where we compare the results from **NIR local** with intermediate outputs at different stages of full energy minimisation. It is seen that the lower part of tea bag under strong shades is gradually recovered with the addition of global reflectance energy.

(a) NIR local (b) Full 150th (c) Full 250th (d) Full 350th (e) Full final

Figure 5: Intrinsics recovered by **NIR local** (5a), and by full energy at different number of iterations (5b) to (5e).

5.4. Convergence and scalability

We plot the energy versus run time curve of our algorithm on the tea bag image scaled to different sizes, in Fig. 6. Our algorithm enjoys a practically linear time complexity, and scales well to images of different sizes. Each curve has a flat tail caused by exhaustive line-search near a minimum. In practice, we may choose to stop the algorithm after a fixed number of iterations (*e.g.* 350, at which point energy has often well-converged) to guarantee a linear complexity implementation. The experiments are carried out on a laptop with 32GB RAM.

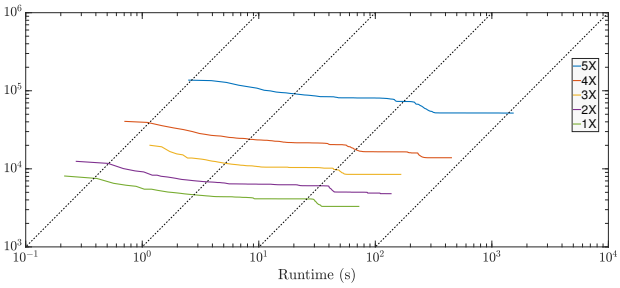


Figure 6: Energy versus runtime for input image of different sizes (size is measured by image width instead of total pixels). Along dotted lines are algorithms of linear time complexities.

6. Conclusion and future work

In this paper an optimisation-based algorithm for NIR-assisted intrinsic decomposition is proposed. We show that the NIR serves as powerful prior that significantly reduces ambiguity. We model this belief with a non-local energy formulation that can be computed and differentiated in linear complexity. In our experiments, the proposed method demonstrates better performance than state-of-arts RGB methods without any training.

Our algorithm requires a NIR-RGB input image pair. While image acquisition is more difficult than RGB-based methods, commodity RGB camera sensors are already able to detect NIR lights, and specialized prism-based NIR-RGB cameras capable of capturing synchronized image pairs have also been commercialized.

We followed the method of [21] to collect ground truth images. In reality we find this set-up difficult to achieve since both coating and realigning objects demand extreme care. We are aiming to expand our dataset by exploring modern 3D printing technologies.

7. Acknowledgement

We would like to thank the reviewers for their valuable comments. This work was finished when Ziang Cheng was visiting National Institute of Informatics, Japan, supported in part by JSPS KAKENHI Grant Number JP15H05918.

References

[1] Andrew Adams, Natasha Gelfand, Jennifer Dolson, and Marc Levoy. Gaussian kd-trees for fast high-dimensional filtering. In *ACM Transactions on Graphics (ToG)*, volume 28, page 21. ACM, 2009. 5

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single RGB-D image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 2

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2015. 1, 2, 6, 7

[4] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978. 1

[5] Anil S Baslamisli, Hoang-An Le, and Theo Gevers. CNN based learning using reflection and retinex models for intrinsic image decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6674–6683, 2018. 2

[6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 1, 4

[7] Nicolas Bonneel, Kalyan Sunkavalli, James Tompkin, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Interactive intrinsic video editing. *ACM Transactions on Graphics (TOG)*, 33(6):197, 2014. 2

[8] Adrien Bousseau, Sylvain Paris, and Frédo Durand. User-assisted intrinsic images. *ACM Transactions on Graphics (TOG)*, 28(5):130, 2009. 2

[9] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 2

[10] Ayan Chakrabarti and Todd Zickler. Statistics of real-world hyperspectral images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 193–200. IEEE, 2011. 1

[11] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[12] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 241–248, 2013. 2

[13] Lechao Cheng, Chengyi Zhang, and Zicheng Liao. Intrinsic image transformation via scale space decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 656–665, 2018. 2

[14] Gyeongmin Choe, Seong-Heum Kim, Sunghoon Im, Joon-Young Lee, Srinivasa G Narasimhan, and In So Kweon. Ranus: RGB and NIR urban scene dataset for deep scene parsing. *IEEE Robotics and Automation Letters*, 3(3):1808–1815, 2018. 1

[15] Gyeongmin Choe, Srinivasa G Narasimhan, and In So Kweon. Simultaneous estimation of near IR BRDF and fine-scale surface geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2452–2460, 2016. 1, 3

[16] Gyeongmin Choe, Jaesik Park, Yu-Wing Tai, and In So Kweon. Exploiting shading cues in kinect IR images for geometry refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3922–3929, 2014. 1

[17] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8944–8952, 2018. 1, 2

[18] Graham D Finlayson, Steven D Hordley, and Mark S Drew. Removing shadows from images using retinex. In *Color and imaging conference*, volume 2002, pages 73–79. Society for Imaging Science and Technology, 2002. 1

[19] Brian V Funt, Mark S Drew, and Michael Brockington. Recovering shading from color images. In *European Conference on Computer Vision*, pages 124–132. Springer, 1992. 2

[20] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. In *Computer graphics forum*, volume 31, pages 1415–1424. Wiley Online Library, 2012. 2

[21] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2335–2342. IEEE, 2009. 1, 2, 4, 6, 7, 8

[22] Qian Huang, Weixin Zhu, Yang Zhao, Linsen Chen, Yao Wang, Tao Yue, and Xun Cao. Multispectral image intrinsic decomposition via subspace constraint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6430–6439, 2018. 1, 2

[23] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*, pages 5936–5946, 2017. 1, 2

[24] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):157, 2011. 1

[25] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *European conference on computer vision*, pages 143–159. Springer, 2016. 2

[26] Ron Kimmel, Michael Elad, Doron Shaked, Renato Keshet, and Irwin Sobel. A variational framework for retinex. *International Journal of computer vision*, 52(1):7–23, 2003. 2

[27] Naejin Kong and Michael J Black. Intrinsic depth: Improving depth transfer with intrinsic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3522, 2015. 1

[28] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. 2

[29] L Lettry, K Vanhoey, and L Van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In *Computer Graphics Forum*, volume 37, pages 409–419. Wiley Online Library, 2018. 2

[30] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. *arXiv preprint arXiv:1808.08601*, 2018. 2

[31] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. *arXiv preprint arXiv:1804.00582*, 2018. 2

[32] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, pages 211–229, 2018. 1, 2

[33] Sosuke Matsui, Takahiro Okabe, Mihoko Shimano, and Yoichi Sato. Image enhancement of low-light scenes with near-infrared flash images. *Information and Media Technologies*, 6(1):202–210, 2011. 1

[34] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2992–2992, 2015. 1

[35] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6789–6798, 2017. 5

[36] Carsten Rother, Martin Kiefel, Lumin Zhang, Bernhard Schölkopf, and Peter V Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. In *Advances in neural information processing systems*, pages 765–773, 2011. 1, 2, 4, 8

[37] Neda Salamati, Clément Fredembach, and Sabine Süsstrunk. Material classification using color and NIR images. In *Color and Imaging Conference*, volume 2009, pages 216–222. Society for Imaging Science and Technology, 2009. 3

[38] Neda Salamati, Diane Larlus, Gabriela Csurka, and Sabine Süsstrunk. Semantic image segmentation using visible and near-infrared channels. In *European Conference on Computer Vision*, pages 461–471. Springer, 2012. 1

[39] Lex Schaul, Clément Fredembach, and Sabine Süsstrunk. Color image dehazing using the near-infrared. In *Proc. IEEE International Conference on Image Processing (ICIP)*, number LCAV-CONF-2009-026, 2009. 1

[40] Ming Shao, Yunhong Wang, and Peijiang Liu. Face relighting based on multi-spectral quotient image and illumination tensorfaces. In *Asian Conference on Computer Vision*, pages 108–117. Springer, 2009. 3

[41] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008. 1, 2

[42] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. 2011. 1, 2, 4

[43] Jian Shi, Yue Dong, Hao Su, and X Yu Stella. Learning non-lambertian object intrinsics across shapenet categories. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5844–5853. IEEE, 2017. 1, 2, 6, 7, 8

[44] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5444–5453. IEEE, 2017. 1

[45] Marshall F Tappen, William T Freeman, and Edward H Adelson. Recovering intrinsic images from a single image. In *Advances in neural information processing systems*, pages 1367–1374, 2003. 1

[46] Ye Yu and William A. P. Smith. Inverserendernet: Learning single image inverse rendering. *CoRR*, abs/1811.12328, 2018. 2

[47] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1437–1444, 2012. 2

[48] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3469–3477, 2015. 2